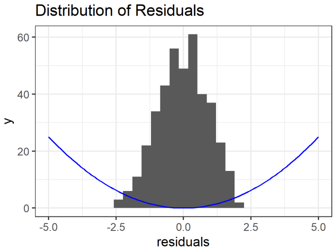
Introduction to Machine Learning

Advanced Risk Minimization
Maximum Likelihood Estimation vs. Empirical Risk Minimization



Learning goals

- Understand the connection between maximum likelihood and risk minimization
- Learn the correspondence between a Gaussian error distribution and the L2 loss

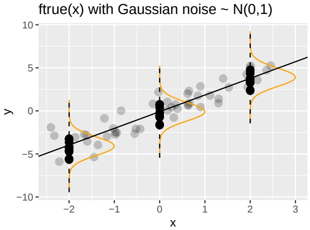
MAXIMUM LIKELIHOOD

Let's consider regression from a maximum likelihood perspective. Assume:

$$y \mid \mathbf{x} \sim p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

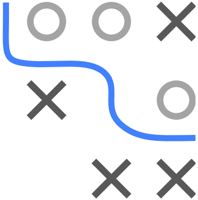
Common case: true underlying relationship f_{true} with additive noise:

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$



where f_{true} has params $\boldsymbol{\theta}$ and ϵ a RV that follows some distribution $\mathbb{P}_\epsilon$, with $\mathbb{E}[\epsilon] = 0$. Also, assume $\epsilon \perp\!\!\!\perp \mathbf{x}$.

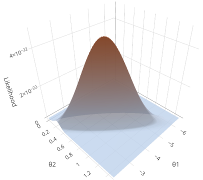
MAXIMUM LIKELIHOOD



From a statistics / maximum-likelihood perspective, we assume (or we pretend) we know the underlying distribution $p(y \mid \mathbf{x}, \boldsymbol{\theta})$.

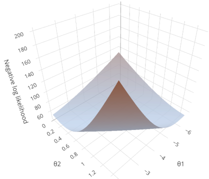
- Then, given i.i.d data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \dots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ from P_{xy} the maximum-likelihood principle is to maximize the **likelihood**

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$



or equivalently to minimize the **negative log-likelihood**

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$



MAXIMUM LIKELIHOOD

From an ML perspective we assume our hypothesis space corresponds to the space of the (parameterized) f_{true}.

- Simply define neg. log-likelihood as **loss function**

$$L\left(y, f(\mathbf{x} \mid \boldsymbol{\theta})\right) := -\log p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

- Then, maximum-likelihood = ERM

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

- NB: When we are only interested in the minimizer, we can ignore multiplicative or additive constants.
- We use $\propto$ as “proportional up to multiplicative and additive constants”

GAUSSIAN ERRORS - L2-LOSS

Assume $y = f_{\text{true}}(\mathbf{x}) + \epsilon$ with additive Gaussian errors, i.e. $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Then

$$y \mid \mathbf{x} \sim N\left(f_{\text{true}}(\mathbf{x}), \sigma^2\right)$$

The likelihood is then

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma^2\right) \\ &\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right) \end{aligned}$$

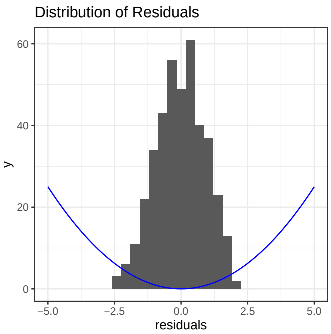
GAUSSIAN ERRORS - L2-LOSS

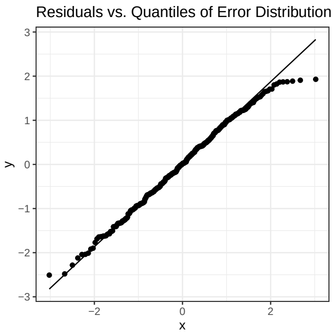
Easy to see: minimizing neg. log-likelihood with Gaussian errors is the same as ERM with $L2$-loss:

$$
\begin{aligned}
-\ell(\boldsymbol{\theta}) &= -\log\left(\mathcal{L}(\boldsymbol{\theta})\right) \\
&\propto -\log\left(\prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right)\right) \\
&\propto \sum_{i=1}^{n}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2
\end{aligned}
$$

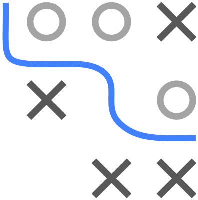
GAUSSIAN ERRORS - L2-LOSS

- We simulate data $y \mid \mathbf{x} \sim \mathcal{N}\left(f_{\text{true}}(\mathbf{x}), 1\right)$ with $f_{\text{true}} = 0.2 \cdot \mathbf{x}$
- Let's plot empirical errors as histogram, after fitting our model with $L2$-loss
- Q-Q-plot compares empirical residuals vs. theoretical quantiles of Gaussian





DISTRIBUTIONS AND LOSSES



- For every error distribution $\mathbb{P}_\epsilon$ we can derive an equivalent loss function, which leads to the same point estimator for the parameter vector $\boldsymbol{\theta}$ as maximum-likelihood. Formally,
  - $\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \implies \hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} -\log(\mathcal{L}(\boldsymbol{\theta}))$
- **But**: The other way around does not always work: We cannot derive a corresponding pdf or error distribution for every loss function – the Hinge loss is one prominent example, for which some probabilistic interpretation is still possible however, see ▸ Sollich 1999.

DISTRIBUTIONS AND LOSSES

When does the reverse direction hold?

- If we can write the loss as $L(y, f(\mathbf{x})) = L(y - f(\mathbf{x})) = L(r)$ for $r \in \mathbb{R}$, then minimizing $L(y - f(\mathbf{x}))$ is equivalent to maximizing a conditional log-likelihood $\log(p(y - f(\mathbf{x}|\boldsymbol{\theta}))$ if
  - $\log(p(r))$ is affine trafo of L (undoing the $\propto$):

$$\log(p(r)) = a - bL(r), \;\; a \in \mathbb{R}, b > 0$$

  - p is a pdf (non-negative and integrates to one)

Thus, a loss L corresponds to MLE under *some* distribution if there exist $a \in \mathbb{R}, \; b > 0$ such that

$$\int_{\mathbb{R}} \exp(a - bL(r)) \mathrm{d}r = 1$$